DOCUMENT RESUME

ED 174 627 TM 008 428

AUTHOR Scheuneman, Janice
TITLE Ethnic Group Bias in Intelligence Test Items.
PUB DATE Mar 78
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Grouping; Black Students; Caucasian Students; Elementary Secondary Education; *Ethnic Groups; Group Intelligence Testing; Intelligence; Intelligence Differences; *Intelligence Tests; Item Analysis; Mexican Americans; Racial Differences; *Racial Discrimination; *Test Bias; *Test Items
IDENTIFIERS Jensen (Arthur R); *Otis Lennon Mental Ability Test

ABSTRACT

In previous studies of ethnic group bias in intelligence test items, the question of bias has been confounded with ability differences between the ethnic group samples compared. The present study is based on a conditional probability model in which an unbiased item is defined as one where the probability of a correct response to an item is the same for all persons of a given ability level regardless of that person's ethnic group background. Using this definition and the chi square procedure for isolating biased items developed by Scheuneman, a large pool of group-administered intelligence test items were examined for ethnic group bias using samples of whites, blacks, and Mexican-Americans. When assessed at grades 8-9, 68 out of 480 items were found to be biased. At grades 11-12, 79 out of 240 items appeared to be biased. In examining the direction of bias, clear evidence of bias against blacks, but no evidence of bias against Mexican-Americans was found. At both levels, the highest proportion of biased items was among the verbal analogy items and letter and figural series completion items. (Author/RD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Janice Scheuneman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED174627

Ethnic Group Bias in Intelligence Test Items*

Janice Scheuneman

The Psychological Corporation

In the controversy over the use of intelligence tests with minority groups, it is usually assumed that such tests are biased in favor of the white middle-class. Despite this commonly held belief, however, relatively little research has been done at the item level to detect differences in response patterns between Blacks and Whites which would suggest that such biases are in fact operating (Loehlin, Lindzey, and Spuhler, 1975). The most comprehensive work which has appeared on this topic to date is that of Jensen (1974, 1976). Jensen distinguishes between the broader question of prediction using intelligence test scores and questions of the internal statistical properties of the test, which he refers to as issues of construct validity. In the 1976 article, he provides a wide variety of evidence supporting the construct validity of several different intelligence tests for Blacks. He first provides evidence that the different tests are measuring the same thing for Blacks and Whites. This evidence includes a comparison of (a) the correlation of raw scores with age, (b) internal consistency reliability indices, and (c) factor loadings of different tests on the first general factor or g. Having thus established that the tests are measuring the same thing, he goes on to compare the performance of Blacks and Whites on the items of the various tests in a number of different ways. His conclusion is that when differences are found, these are consistent with the hypothesis that the samples being examined differ in the ability being measured. Hence he discounts the likelihood that cultural bias is operating as well.

The weight of Jensen's evidence is compelling. He has reported several different studies relating to this issue all with similar results. However,

*Paper presented at the meeting of the American Educational Research Association, Toronto, March 1978. Printed in U.S.A

TM008 428

there are two major points which should be noted. First, while the evidence that the samples being studied differ in ability is difficult to dispute, Jensen does not provide evidence to suggest that the differences are other than would be expected quite apart from racial considerations. A number of factors other than ethnic group membership are known to affect IQ scores, and the samples used in Jensen's studies are apt to differ in a variety of these other factors. Hence the composition of the samples compared is crucial in evaluating this work. Unfortunately, Jensen does not describe his samples in any detail. In the 1976 paper, he reports only that "representative" samples were used. The question immediately arises, representative in what way? Samples representative of the ethnic group populations of which they are a part would surely not be expected to have equivalent performance on intelligence tests for reasons other than ethnicity. In the 1974 article, he does specify that the samples were representative of the ethnic group populations in the particular community from which they were drawn, a community which was relatively homogeneous with regard to SES.

Trotman (1977) produces evidence to suggest, however, that even if samples are in fact equivalent on SES, they may not be equivalent on a number of home environmental factors known to affect intellectual development. In her investigation, she obtained home environment indices for 50 White and 50 Black middle-class girls. The two samples were found to be very similar in SES according to Warner's Index of Status Characteristics and yet still exhibited large and significant differences in the intellectual home environment index obtained through interviews with each girl's mother. The home environment index correlated .68 with IQ score for Blacks and .37 for Whites. Clearly such factors are related to obtained scores, even within the relatively restricted SES range.

The second point to be considered relates to the use of samples differing in ability when studying item bias. Let us suppose that, while true differences in ability exist between samples drawn from different ethnic groups, additional factors relating to ethnic background are also operating to further suppress the scores of the minority group. Little if any of the evidence presented by Jensen is inconsistent with such a possibility. In fact when Jensen (1976) reports testing for item by group interaction, generally accepted as one means of assessing this possibility, the interactions were found to be significant. Jensen discounts this finding, however, because the variance accounted for was small compared to that produced by ability differences. It should be noted, however, that to the extent such an effect is consistent across items, it would be attributed to differences in ability rather than to the interaction.

It would be very difficult if not impossible to find samples which are equivalent in every important way except ethnic group membership in order to compare their performance on intelligence tests appropriately. However, it is possible, using the samples available, to turn the problem around and compare persons of different ethnic background, but roughly equivalent ability in their performance on these tests. Within such a context, an item which is not biased would be one for which the probability of a correct response is the same for all persons of a given ability level regardless of ethnic group background. With this model, the distribution of ability within the different ethnic groups becomes unimportant. The study reported in this paper takes this approach to study the performance of samples from three ethnic groups--Blacks, Whites, and Mexican-Americans--on a large pool of fairly heterogeneous intelligence test items. The purpose of this study is to determine if bias exists in these items and, if so, what kinds of items appear to be biased most often.

Method

This study was conducted as part of the item analysis program for the 1979 revision of the Otis-Lennon Mental Ability Test. Data were taken from those sites participating in the item analysis program that had a minority representation of 15% or more within the community. Two additional sites with over 50% minority population were also tested especially for the bias study in order to assure adequate sample sizes for the two minority groups. The White sample was made up of those tested in the same sites as the minorities. All testing sites were located in the south or southwest. Ethnic group membership was identified by the examinee.

The item pool was contained in six booklets for the Intermediate Level (grades 4-9) and three booklets for the Advanced level (grades 10-12). Although all items were intended to be a measure of general intelligence or g, the items were divided into four subscales for test construction purposes and each of these subscales was scored separately. Each booklet contained 15 Verbal Comprehension items, 30 Verbal Reasoning, 20 Quantitative Reasoning, and 15 Figural Reasoning items. Thus, with 80 items per booklet, the total pool at the Intermediate Level was 480 items and at the Advanced Level, 240 items. All booklets were arranged in a spiral omnibus format roughly in order of difficulty and were administered without time limit.

The procedure used for evaluating the items for bias was Scheuneman's chi square technique (1975, 1976). In this method, an unbiased item is defined as above, and ability is measured by total score on the test or on a subset of items containing the item in question. The ability distribution is then divided into a small number of subgroups each represented by a range of total scores, such that the ability groups are mutually exclusive and exhaustive. The number of correct responses made by

members of each ethnic group within each ability range is the obtained frequency used in the computation of the chi square. The expected frequencies are then determined from the proportion of correct responses of all subjects scoring within an ability range regardless of ethnic group membership. This procedure has been evaluated and compared with other procedures for assessing bias in test items by Rudner (1977) and Nungester (1977). These studies show that results using this procedure agree well with those obtained with other techniques. In this study, the four subscales--Verbal Comprehension, Verbal Reasoning, Quantitative Reasoning, and Figural Reasoning--were used as ability measures as these were more homogeneous than the test total.

The Intermediate level items were evaluated for bias at grade 5 and again at grades 8-9. The Advanced level items were evaluated only once at grades 11-12. Nearly half the items at grade 5, however, had too few correct responses to be analyzed for bias. Since this made the patterns of results ambiguous, only the results for grades 8-9 and grades 11-12 will be reported here. The six Intermediate level booklets at grades 8-9 and the three Advanced level booklets at grades 11-12 were randomly assigned to pupils within a classroom. Samples taking the different booklets at a given grade level should therefore be comparable. At grades 8-9, sample sizes for the six booklets ranged from 218 to 254 for Whites, 181 to 209 for Blacks, and 132 to 150 for Mexican-Americans. Samples for the three booklets at grades 11-12 were comparable, ranging in size from 294 to 312 for Whites, 162 to 174 for Blacks, and 94 to 102 for Mexican-Americans.

Results

All items were analyzed for bias using three to five ability ranges where the subscale score was the ability measure. The chi square values obtained,

however, were smaller than would be expected by chance for the appropriate degrees of freedom. (Degrees of freedom with this test are (r-1)(k-1) where r is the number of ability range groups and k is the number of ethnic groups.) Hence, although the probabilities of the chi square values were read using a standard table in order to compare results which had been obtained using different numbers of ability intervals, these probabilities will not correspond to the probability of obtaining that chi square value when the null hypothesis is true. The level selected as a cutoff point was a probability of .30, with items exceeding the usual .05 examined separately.

When evaluated using this procedure, 66 items at grades 8-9 and 79 items at grades 11-12 had chi square values sufficiently high to meet the criterion for bias. These items represent 14 and 33 percent of their respective item pools. However, since the criterion was arbitrarily set and the true alpha level is unknown, it is important to establish both that the results are meaningful and tend to be consistent across levels, and that they are different than would be expected if items had been selected randomly as biased. Results will be examined from two perspectives, (a) the apparent direction of the bias for the three groups and (b) the types of items which tend to be biased at the two levels.

When the procedure is used with just two groups, results can only indicate that one group appears to be favored over the other. With three groups, however, it is possible to determine if one or more of the groups is performing about as expected while another is either performing better or less well. The direction of bias can be inferred by comparing the expected and obtained frequencies. For example, when the obtained number of correct responses for an ethnic group is less than expected in all or most of the ability ranges, the item would be considered biased against that group. When the chi square is

sufficiently large to meet the criterion for bias, that part of the chi square which came from each of the ethnic groups can be determined to suggest which of the groups is contributing to the significant result. In Table 1, an item is illustrated which appears to be both biased against Blacks and in favor of Whites. Notice that the sum of the cell square contingencies (Totals for Groups) for both Whites and Blacks is around nine, while for Mexican-Americans it is close to zero. Table 2 shows an item which also appears to be biased against Blacks, but performance of Whites and Mexican-Americans is only slightly better than would be expected. The totals for groups with this item are 8.03 for Blacks, 3.23 for Whites, and 2.24 for Mexican-Americans.

The direction of bias for the items studied is given in Table 3. In tabulating items for bias in either direction, a total of at least three for that ethnic group was arbitrarily selected as a criterion. Some items were counted more than once, since, for example, an item could be biased in favor of one group and against the other two. Items for which the direction was inconsistent across ability ranges or where the group total was less than three for all three of the ethnic groups were not included in this tabulation.

The results in Table 3 show a clear pattern of bias against Blacks. Except for the most biased items at grades 8-9, where the frequencies were too small to test, the number of items found to be biased against or in favor of the different groups was significantly different than would be expected by chance. (Using a 2 x 3 contingency table and the normal chi square test, chi square values obtained were 23.50 for Grades 8-9, and 50.48 and 35.46 at Grades 11-12 for all biased items and the most biased items respectively. At two degrees of freedom, all are significant well beyond .01.) On the other hand, the evidence here does not support a hypothesis of bias agianst Mexican-Americans. At grades 8-9, the number of biased items is not large, and the direction of

bias is about as expected from the overall proportions of items in each direction. At grades 11-12, however, the pattern is much like that for Whites in both the number and direction of biased items. It is possible that in some way this was a select sample. A large proportion of the Mexican-American sample was enrolled in parochial schools, which may make them atypical. When asked about bilingual status, 47 to 71 percent of those responding on the different booklets at the Intermediate Level and 74 to 81 percent of those at the Advanced Level reported that they were bilingual.

Different types of items appeared within each of the four subscales on the tests. In all, the Intermediate level tests contained 16 different item types and the Advanced tests contained 14. Tests at the two levels contained different items, however, with very few exceptions. Table 4 gives the distribution of biased items according to the different item types. At grades 8-9, five item types--verbal analogies, figure series, letter series, word matrices, and opposites--had more than 20 percent biased items. The lowest were scrambled sentences and number matrices, with no biased items, and number series, with only five percent. At grades 11-12, verbal analogies, figure series, and letter series were again among the item types with the highest percent of bias. Sentence completion and number problems also had very high percentages of biased items at this level. The lowest percents were with the straightforward number manipulation items--number series, number pairs, and number matrices. Also quite low percents of biased items were found in verbal classifications and figure analogies. Both of these were also quite low at grades 8-9. In general, the tendency is for the same types of items to be biased most often and least often at the two levels. Although frequencies are too small for significance tests, it appears clear that the obtained patterns are unlikely to have occurred by chance.

Within the three item categories that are high for both levels, the direction for the different groups was also examined. With the analogies, the most prevalent pattern appeared to be against Blacks. At grades 8-9 and grades 11-12, eleven and nine items respectively were found to be biased against Blacks. The percentage goes up when examining only the most biased items, where seven out of eight at grades 8-9 and five out of six at grades 11-12 were biased against Blacks. The figure series items appeared to favor Whites at Intermediate and to be biased against Blacks at Advanced. Letter series on the other hand seemed to favor Mexican-Americans. The number problems at the Advanced level are also noteworthy. Five of the 12 items in the pool were found to be biased beyond the more stringent criteria. All were biased against Blacks.

It seems clear that some number of items must have been selected as biased by chance so that it is easy to dismiss those items in categories with low frequencies of bias as being among the chance results. However, some surprising consistencies occur even here. For example, among the five number series items found to be biased at one level or the other, two required that a number other than the last in the series be filled in and two depended more on the pattern made by the numbers than on the quantitative relationships among them. Both of these types of series were infrequently included in the test. Another interesting example comes from the classification items. One of the two biased items at the Advanced level included the terms, noisy, fragrant, bitter, rough and thoughtless. At the Intermediate level, one of the three biased items included sour, bitter, sweet, salty, and spoiled. Both items were thus conceptually similar and both were biased in favor of Blacks. Two other items found to be biased in favor of Blacks at the Intermediate Level are inference items, both of which involve brothers--"If a man has brothers, but not nephews..." and "A few of the people who know my brother,

know me..." Perhaps in a gross way, these items sound more relevant to Black children than do some of the other items. Hence, the results appear to support the contention that at least in large part the items found to be biased reflect genuine response differences among the three ethnic groups.

Discussion and Conclusions

This study has used a large sample of intelligence test items taken from the item analysis edition of a group-administered intelligence test. These items were evaluated for bias using a method which compares the performance of persons of roughly equivalent ability, rather than the performance of groups with differing distributions of ability. With this technique, clear evidence of bias against Blacks in intelligence test items was found. Contrary to expectation, however, no evidence of bias against Mexican-Americans was found using this approach.

In addition to the tabulated results presented here, certain observations were made which suggest possibilities for further research. It was noted, for example, that no sample item was given which explained what was required for either letter series or figural series items. The only instructions were given in the item stem. Any uncertainty the pupil may have felt in responding to such an item might be reflected in just the sort of response decrement shown in the bias results. Consequently, an example of each of these item types was added to the final version of the test, thus increasing the number of samples from three to five.

Another observation concerned the proportion of biased items, which at grades 11-12 was more than double that at grades 8-9. It is interesting to speculate on why this should have occurred. Sample size is known to effect the size of the chi square when using this technique, but sample sizes were quite comparable for the different booklets. Easy items have generally

been found to be biased less often than difficult items, but again, the Advanced level items were only slightly more difficult on the average than were the Intermediate items at the appropriate grade levels. These results may, however, reflect a genuine difference in the tests. At least one possible explanation lies in the type of subject matter required to achieve the item difficulty needed for effective measurement. Although the results for grade 5 were not clear in many ways, few of the verbal analogy items which were tested at that level appeared to be biased, a departure from the results at the two higher levels. At grade 5, most of the analogies tested involved body parts, common foodstuffs, and other items which should be common in diverse cultures. However, as the level increases, such items no longer have adequate difficulty. By high school, those items which proved to be sufficiently difficult were those with the least common vocabulary and concepts. It may not be possible to build difficult items where the probability of previous exposure is the same for persons of diverse background. At least, such a task poses a challenge for those of us in the measurement community concerned with the construction of these tests.

In building the revised Otis-Lennon, in addition to the change in the examples, those items which appeared to be biased using this procedure were not used whenever possible. However, it is not clear what effect these efforts will actually have on scores. Although this study supports the hypothesis that bias is a factor for Blacks, it provides no evidence concerning the impact of such bias on results nor can it estimate how much, if any, of the observed differences between Blacks and Whites can be accounted for by such items. In the aggregate, the effect of biased items may be rather small, yet the impact on an individual pupil's score may still be significant.

The use of the item analysis version of the intelligence test made

possible the examination of a large number of items defined more specifically than verbal and non-verbal. The difference in the results for different types of verbal items suggests that this dichotomy is an oversimplification in discussing test bias. For example, while verbal analogies are frequently biased at both levels, verbal classification items are very infrequently biased at either level. These kinds of results, as well as those of Trotman concerning the inadequacy of SES as a controlling variable of home background, suggest that the problem of bias in intelligence testing is far more subtle and complex than either the critics or defenders of these tests appear to believe. More research on the nature and effects of biases in intelligence tests is required if we are to use them intelligently with minority group persons.

References

Jensen, A. R. How biased are culturally loaded tests? Genetic Psychology Monographs, 1974, 90, 185-244.

Jensen, A. R. Test bias and construct validity. Phi Delta Kappan, 1976, 58, 340-346.

Loehlin, J. C., Lindzey, G., and Spuhler, J. N. Race differences in intelligence. San Francisco: W. H. Freeman, 1975.

Nungester, R. J. An empirical examination of three models of item bias. Unpublished doctoral dissertation, The Florida State University, 1977.

Rudner, L. M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University of America, 1977.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the meeting of American Educational Research Association, Washington, April 1975.

Scheuneman, J. Validating a procedure for assessing bias in test items in the absence of an outside criterion. In W. Coffman (Chair), The assessment of bias in the absence of an outside criterion. Symposium presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

Trotman, F. K. Race, IQ, and the middle class. Journal of Educational Psychology, 1977, 69, 266-273.

Table 1

Computation of the Chi Square

for one Item with Three Groups

Range	Group	No. Scoring in interval[a]	No. Correct	Expected Proportion[b]	Expected no. correct[c]	Cell square contingencies[d]
16-20	White	89	72		71.05	.01
	Black	14	11		11.18	.00
	Mex.-Am.	16	12		12.77	.05
	Total	119	95	.798		
11-15	White	106	67		59.81	.86
	Black	35	13		19.75	2.31
	Mex.-Am.	38	21		21.44	.01
	Total	179	101	.564		
0-10	White	69	25		13.86	8.94
	Black	110	10		22.10	6.63
	Mex.-Am.	35	8		7.03	.13
	Total	214	43	.201		
Total Chi square						18.95

Totals for Groups

Whites 9.82 Blacks 8.94 Mex.-Am. 0.19

[a]This figure has been adjusted by omitting those persons who did not respond to the item. A total of 58 persons omitted this item.

[b]This number is obtained by dividing the total number of correct responses within the interval by the total number scoring in the interval.

[c]This number is obtained by multiplying the expected proportion correct by the total number in each ethnic group scoring in that interval.

[d]This number is the usual $(fo-fe)^2/fe$ where fo is the number of correct responses and fe is the expected number of correct responses for each ethnic group.

Table 2

Computation of the Chi Square

for one Item with Three Groups

Range	Group	No. Scoring in interval[a]	No. correct	Expected Proportion[b]	Expected no. correct[c]	Cell square contingencies[d]
23-30	White	69	47		45.59	.04
	Black	14	7		9.25	.55
	Mex.-Am.	29	20		19.16	.04
	Total	112	74	.661		
15-22	White	109	37		33.47	.37
	Black	65	10		19.96	4.97
	Mex.-Am.	17	27		20.57	2.01
	Total	241	74	.307		
0-14	White	55	14		8.97	2.82
	Black	93	9		15.17	2.51
	Mex.-Am.	42	8		6.85	.19
	Total	190	31	.163		
Total Chi Square						13.50

Totals for Groups

Whites 3.23 Blacks 8.03 Mex.-Am. 2.24

[a]This figure has been adjusted by omitting those persons who did not respond to the item. A total of 22 persons omitted this item.

[b]This number is obtained by dividing the total number of correct responses within the interval by the total number scoring in the interval.

[c]This number is obtained by multiplying the expected proportion correct by the total number in each ethnic group scoring in that interval.

[d]This number is the usual $(fo-fe)^2/fe$ where fo is the number of correct responses and fe is the expected number of correct responses for each ethnic group.

Table 3

Apparent Direction of Bias in items

	Grades 8-9		Grades 11-12	
	In favor	Against	In favor	Against
All biased items[a]				
Whites	26	6	15	3
Blacks	7	24	1	31
Mex.-Am.	13	6	28	6
Total	46	36	44	40
Most biased items[b]				
Whites	10	1	13	3
Blacks	1	12	1	24
Mex.-Am.	1	2	13	2
Total	12	15	27	29
Number of items:				
All biased	58		61	
Most biased	14		29	
Not Tabulated	10		18	

[a]Probability of obtained chi squares is less than .30.

[b]Probability of obtained chi squares is less than .05.

Table 4

Number of Biased Items by Item Type

Item type	Grades 8-9					Grades 11-12				
	No. in pool	All biased items		Most biased items		No. in pool	All biased items		Most biased items	
		No.	%	No.	%		No.	%	No.	%
Verbal Comprehension										
Opposites	42	9	21	1	2	24	6	25	2	8
Sentence Completion	36	3	8	0		21	11	52	5	24
Scrambled Sentences	12	0		0						
Verbal Reasoning										
Inference	30	5	17	2	7	18	5	28	1	6
Analogies	66	18	27	8	12	39	20	51	6	15
Classification	36	3	8	0		21	2	9	2	9
Word Matrix	24	5	21	1	4	12	4	33	1	8
Logical Selection	24	2	8	0						
Quantitative Reasoning										
Letter Series	24	5	21	1	4	12	7	58	2	17
Problems	18	3	17	1	6	12	6	50	5	42
Number Series	42	2	5	0		18	3	17	0	
Number Pairs	24	2	8	0		12	1	8	0	
Number Matrix	12	0		0		6	1	17	0	
Figural Reasoning										
Series	30	7	23	0		15	6	40	3	20
Analogies	36	2	6	0		18	3	17	2	11
Matrix	24	2	8	0		12	4	33	0	
Total	480	68	14	14	3	240	79	33	29	12